CAR-TR-799
CS-TR-3555

N00014-95-1-0521
November 1995

3D Model-Based Tracking of Humans in Action: A Multi-View Approach

D.M. Gavrila
L.S. Davis

Computer Vision Laboratory
Center for Automation Research
University of Maryland
College Park, MD 20742-3275

http://www.umiacs.umd.edu/users/{gavrila,lsd}/

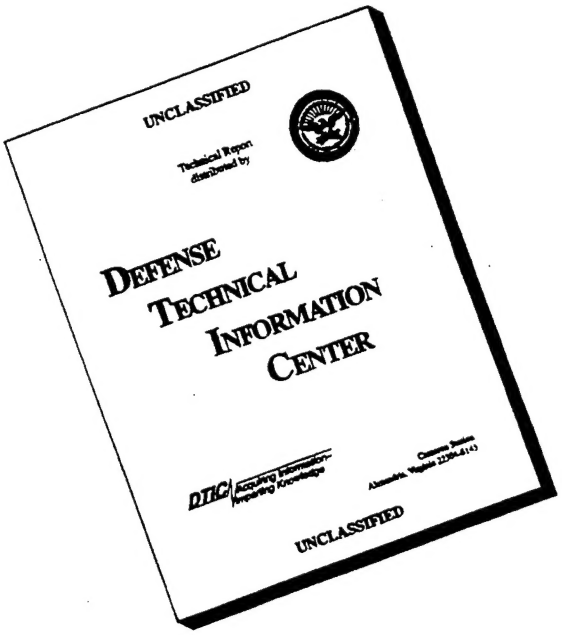
Abstract

We present a vision system for the 3D model-based tracking of unconstrained human movement. Using image sequences acquired simultaneously from multiple views, we recover the 3D body pose at each time instant without the use of markers. The pose-recovery problem is formulated as a search problem and entails finding the pose parameters of a graphical human model whose synthesized appearance is most similar to the actual appearance of the real human in the multi-view images. The models used for this purpose are acquired from the images. We use a decomposition approach and a best-first technique to search through the high dimensional pose parameter space. A robust variant of chamfer matching is used as a fast similarity measure between synthesized and real edge images.

We present initial tracking results from a large new Humans-In-Action (HIA) database containing more than 2500 frames in each of four orthogonal views. The four image streams are synchronized. They contain subjects involved in a variety of activities, of various degrees of complexity, ranging from simple one-person hand waving to two-person close interaction in the Argentine tango.

The support of the Advanced Research Projects Agency (ARPA Order No. C635) and the Office of Naval Research under Grant N00014-95-1-0521 is gratefully acknowledged, as is the help of Sandy German in preparing this paper.

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

1 Introduction

The ability to recognize humans and their activities by vision is a key feature in the pursuit of designing machines capable of interacting intelligently and effortlessly in a human-inhabited environment. Besides this long-term goal, many applications are possible in the relatively near term, e.g. in virtual reality, "smart" surveillance systems, motion analysis in sports, choreography of dance and ballet, sign language translation, and gesture-driven user interfaces. In many of these applications a non-intrusive sensory method based on vision is preferable over a method (in some cases not even feasible) that relies on markers attached to the bodies of human subjects.

Our approach to looking at humans and recognizing their activities has two major components:

1. body pose recovery and tracking
2. recognition of movement patterns

Several choices have to be made in connection with body pose determination and tracking, which affect what features can be used: the type of model used (stick figure, volumetric model, none), the dimensionality of the space in which tracking takes place (2D or 3D), the number of sensors used (single, stereo, multiple), the sensor modality (visible light, infrared, range), the sensor placement (centralized vs. distributed) and mobility (stationary vs. moving). We consider the case where we have multiple stationary (visible-light) cameras, previously calibrated, and we observe one or more humans performing actions from multiple viewpoints. The aim of the first component of our approach is to reconstruct from the sequence of multi-view frames the (approximate) 3D body pose(s) of the human(s) at each time instant; this serves as input to the movement recognition component. In an earlier paper [6] movement recognition was considered as a classification problem and a Dynamic Time Warping method was used to match a test sequence with several reference sequences representing prototypical activities. The features used for matching were various 3D joint angles of the human body. In this paper, we focus on the pose recovery and tracking component of our system.

The outline of this paper is as follows. Section 2 provides a motivation for our choice of a 3D recovery approach rather than a 2D approach. In Section 3 we discuss 3D human modeling issues and the (semi-automatic) model acquisition procedure used by our system. Section 4 deals with the pose recovery and tracking component. Included is a bootstrapping procedure to start the tracking or to re-initialize it if it fails. Section 5 presents new experimental results in which successful unconstrained whole-body movement is demonstrated on two subjects. These are initial results[1] derived from a large Humans-In-Action (HIA) database containing two subjects involved in a variety of activities, of various degree of complexity. We discuss our results and possible improvements in Section 6. Finally, Section 7 contains our conclusions.

2 2D vs. 3D

One may question whether it is desirable or feasible to try to recover 3D body pose from 2D image sequences for the purpose of recognizing human movement. An alternative approach is to work directly with 2D features derived from the images. Model-free 2D features are usually obtained by applying a motion-detection algorithm to the image (assuming a stationary camera) and obtaining the outline of a moving object, presumably human. Frequently, a $K \times N$ spatial grid is superimposed on the motion region, after a possible normalization of its extent. In each of the $K \times N$ tiles a simple feature is computed, and these are combined to form a $K \times N$ feature vector to describe the state of movement at time t. This is the approach taken by Polana and Nelson [23] and Darrell and Pentland [4]. Another possibility is to use 2D model-based features, where the assumption is that as a result of 2D segmentation and tracking a sequence of 2D stick figure poses is available. For example, Goddard [8] uses the 2D angular velocities and orientations of the links as features. Guo et al. [10] uses a combination of link orientations and joint positions of the stick figure.

Recognition systems using 2D model-free features have had early successes in matching human movement patterns. For constrained types of human movement (such as walking parallel to the image plane, involving periodic motion), many of these features have been

[1]The tracking results described in this paper are also available as video clips from our home pages.

successfully used for classification, as in [23]. This may indeed be the easiest and best solution for several applications. But we find it unlikely that reliable recognition of more unconstrained and complex human movements (e.g. humans wandering around, making gestures while walking and turning) can be achieved using these types of features exclusively. With respect to using 2D model-based features, we note that few systems actually derive the features they use for movement matching. Self-occlusion makes the 2D tracking problem hard for arbitrary movements and thus existing systems assume some a priori knowledge of the type of movement and/or the viewpoint under which it is observed [1, 19]. 2D labeling and tracking under more general conditions is attempted by [16].

We therefore investigate in this paper the more general-purpose approach of recovering 3D pose through time, in terms of 3D joint angles defined with respect to a human-centered [17] coordinate system. 3D motion recovery from 2D images is often an ill-posed problem. In the case of 3D pose tracking, however, we can take advantage of the available a priori knowledge about the kinematic and shape properties of the human body to make the problem tractable. Tracking also is well supported by the use of a 3D human model which can predict events such as (self) occlusion and (self) collision. Once 3D tracking is successfully completed, we have the benefit of being able to use the 3D joint angles as features for movement matching, which are viewpoint independent and directly linked to the body pose. Compared with 3D joint coordinates, they are less sensitive to variations in the size of the human.

The techniques described in this paper lead to tracking on a fine scale, with the obtained joint angles being within a few degrees of their true values. Besides providing meaningful generic features for a movement matching component, such techniques are of independent interest for their use in virtual reality applications. In other applications, such as surveillance, continuous fine-scale 3D tracking will not always be necessary, and can be combined with tracking on a more coarse level (for example, considering the human body as a single unit), changing the mode of operation from one to another depending on context. For related work by Intille and Bobick see [13].

3 3D body modeling and model acquisition

3D graphical models for the human body generally consist of two components: a representation for the skeletal structure (the "stick figure") and a representation for the flesh surrounding it. The stick figure is simply a collection of segments and joint angles with various degree of freedom at the articulation sites. The representation for the flesh can either be surface-based (using polygons, for example) or volumetric (using cylinders, for example). There is a trade-off between the accuracy of representation and the number of parameters used in the model. Many highly accurate surface models have been used in the field of graphics [2] to model the human body, often using thousands of polygons obtained from actual body scans. In vision, where the inverse problem of recovering the 3D model from the images is much harder and less accurate, the use of volumetric primitives has been preferred to "flesh out" the segments because of the lower number of model parameters involved.

For our purposes of tracking 3D whole-body motion, we currently use a 22-DOF model (3 DOF for the positioning of the root of the articulated structure, 3 DOF for the torso and 4 DOF for each arm and each leg), without modeling the palm of the hand or the foot, and using a rigid head-torso approximation. See [2] for more sophisticated methods of modeling. Regarding shape, we felt that simple cylindrical primitives (possibly with elliptic XY-cross-sections) [5, 11, 25] would not represent body parts such as the head and torso accurately enough. Therefore, we employ the class of *tapered super-quadrics* [18]; these include such diverse shapes as cylinders, spheres, ellipsoids and hyper-rectangles. Their parametric equation $\mathbf{e} = (e_1 e_2 e_3)$ is given by [18]

$$\mathbf{e} = a \begin{pmatrix} a_1 C_u^{\epsilon_1} C_v^{\epsilon_2} \\ a_2 C_u^{\epsilon_1} S_v^{\epsilon_2} \\ a_3 S_u^{\epsilon_1} \end{pmatrix} \tag{1}$$

where $-\pi/2 \leq u \leq \pi/2, -\pi \leq v \leq \pi$, and where $S_\theta^\epsilon = \text{sign}(\sin\theta)|\sin\theta|^\epsilon$, and $C_\theta^\epsilon = \text{sign}(\cos\theta)|\cos\theta|^\epsilon$. In (1), $a \geq 0$ is a scale parameter, $a_1, a_2, a_3 \geq 0$ are aspect ratio parameters, and ϵ_1, ϵ_2 are "squareness" parameters. Adding linear tapering along the z-axis to the

super-quadric leads to the parametric equation $s = (s_1 s_2 s_3)$ [18]:

$$\mathbf{s} = \begin{pmatrix} \left(\frac{t_1 e_3}{a a_3} + 1\right) e_1 \\ \left(\frac{t_2 e_3}{a a_3} + 1\right) e_2 \\ e_3 \end{pmatrix} \tag{2}$$

where $-1 \leq t_1, t_2 \leq 1$ are the taper parameters along the x and y axes. So far, we have obtained satisfactory modeling results with these primitives alone (see experiments); a more general approach also allows deformations of the shape primitives [18, 21].

In this work, we derive shape parameters $\mathbf{S_k} = (a^k, a_1^k, a_2^k, a_3^k, \epsilon_1^k, \epsilon_2^k, t_1^k, t_2^k)$ from the projections of occluding contours in two orthogonal views, parallel to the zx- and zy-planes. This involves the human subject facing the camera frontally and sideways. We assume 2D segmentation of the two orthogonal views; a way to obtain such a segmentation is proposed in recent work by Kakadiaris and Metaxas [15]. Back-projecting the 2D projected contours of a quadric gives the 3D occluding contours, after which a coarse-to-fine search procedure is used over a reasonable range of parameter space to determine the best-fitting quadric. Fitting uses chamfer matching (see the next section) as a similarity measure between the fitted and back-projected occluding 3D contours. Figure 1 shows frontal and side views of the recovered torso and head for two persons: DARIU and ELLEN. Figure 2 shows their complete recovered models in a graphics rendering. These models are used in the tracking experiments of Section 5.

4 Pose recovery and tracking

The general framework for our tracking component is adapted from the early work by Rourke and Badler [26] and is illustrated in Figure 3a. Four main components are involved: prediction, synthesis, image analysis and state estimation. The prediction component takes into account previous states up to time t to make a prediction for time $t + 1$. It is deemed more stable to do the prediction at a high level (in state space) than at a low level (in image space), allowing an easier way to incorporate semantic knowledge into the tracking process. The synthesis component translates the prediction from the state level to the measurement (image) level, which allows the image analysis component to selectively focus on a subset of

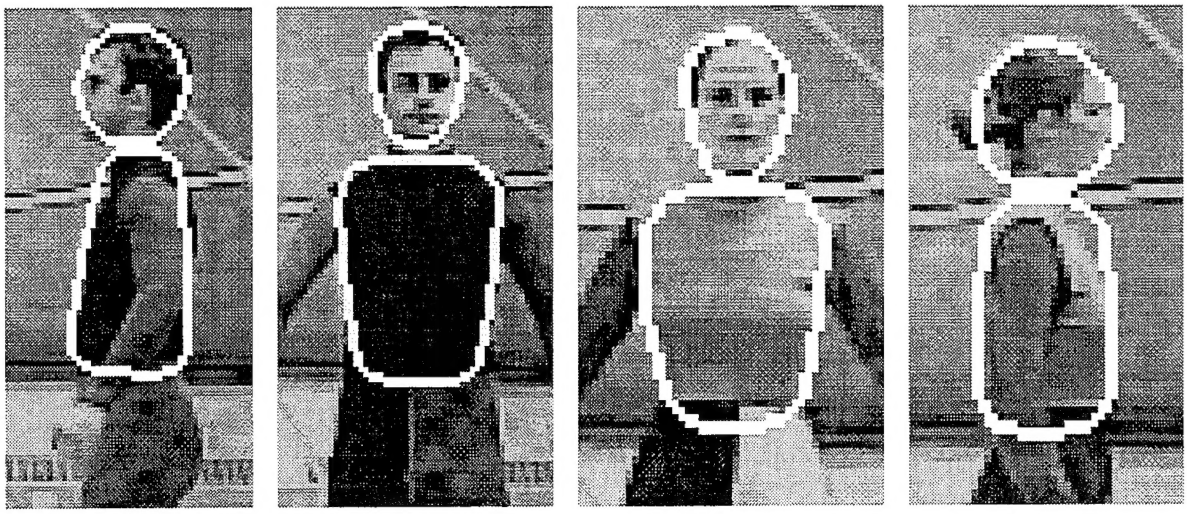


Figure 1: Frontal and side views of the recovered torso and head for the DARIU and ELLEN models.

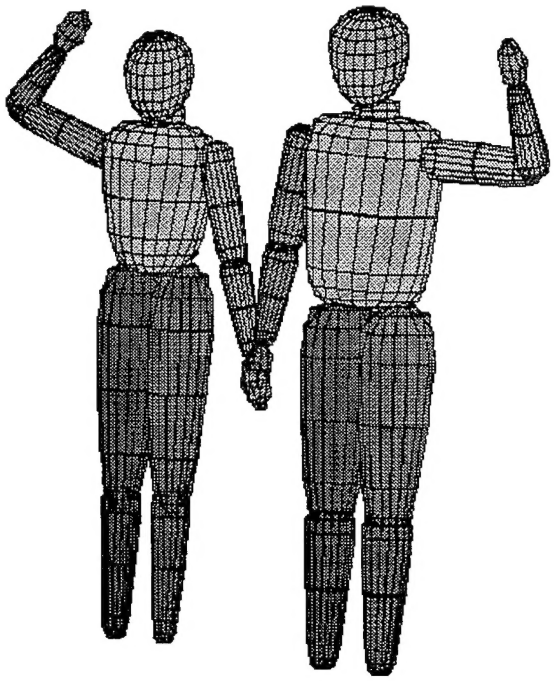


Figure 2: The recovered 3D models ELLEN and DARIU say "hi!"

regions and look for a subset of features. Finally, the state-estimation component computes the new state using the segmented image.

The above framework is general and can also be applied to other model-based tracking problems. In the remainder of this section, we discuss how the components are implemented in our system for the case of tracking humans, and how this relates to existing work. In

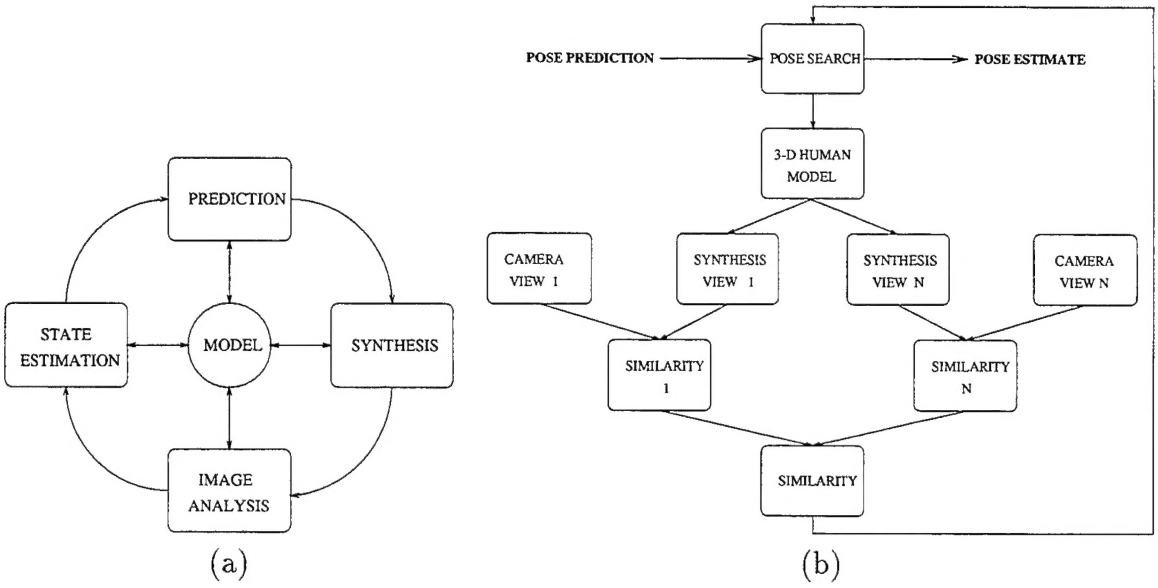


Figure 3: (a) Tracking cycle; (b) pose-search cycle.

the first subsection we cover the pose estimation component; the second subsection briefly covers the other components.

4.1 Pose estimation

One approach to pose recovery is to derive point matches between a 3D figure and its 2D projection to solve for the former, perhaps using several images. The advantage of this is that rigorous mathematical analysis can be applied to solve for the 3D pose; the problem can be solved using techniques borrowed from inverse kinematics (see the precursor to [24]), constrained optimization [29], or algebraic geometry [12]. On the downside, this approach requires feature points (usually the joints) to be accurately located in the images, which is quite difficult. Moreover, the approach seems to be very sensitive to occlusion.

We therefore pursued an alternative approach to pose recovery, based on a generate-and-test strategy. Here, the pose recovery problem is formulated as a search problem and entails finding the pose parameters of a graphical human model whose synthesized appearance is most similar to the actual appearance of the real human (see Figure 3b). This approach has the advantage that the measure of similarity between synthesized appearance and actual appearance can now be based on whole contours and/or regions rather than on a few points.

So far, existing systems which work on real images using this strategy have had limitations. Perales and Torres [22] describe a system which involves input from a human operator. Hogg [11] and Rohr [25] deal with the restricted movement of walking parallel to the image plane, for which the search space is essentially one-dimensional. Downton and Drouet [5] attempt to track unconstrained upper-body motion, but conclude that the tracking fails due to propagation of errors. Recent work by Goncalves et al. [9] uses a Kalman-filtering approach to track arm movements from single-view images where the shoulder remains fixed. Finally, work by Rehg [24] is geared towards finger tracking. We aim to improve the previous approaches, where applicable, along the following lines.

Similarity measure

In our approach the similarity measure between model view and actual scene is based on arbitrary edge contours rather than on straight line approximations (as in [25], for example); we use a robust variant of *chamfer matching* [3]. The *directed* chamfer distance $DD(T, R)$ between a test point set T and a reference point set R is obtained by summing the distances between each point in set T to its nearest point in R:

$$DD(T, R) = \sum_{t \in T} dd(t, R) = \sum_{t \in T} \min_{r \in R} \| t - r \| \tag{3}$$

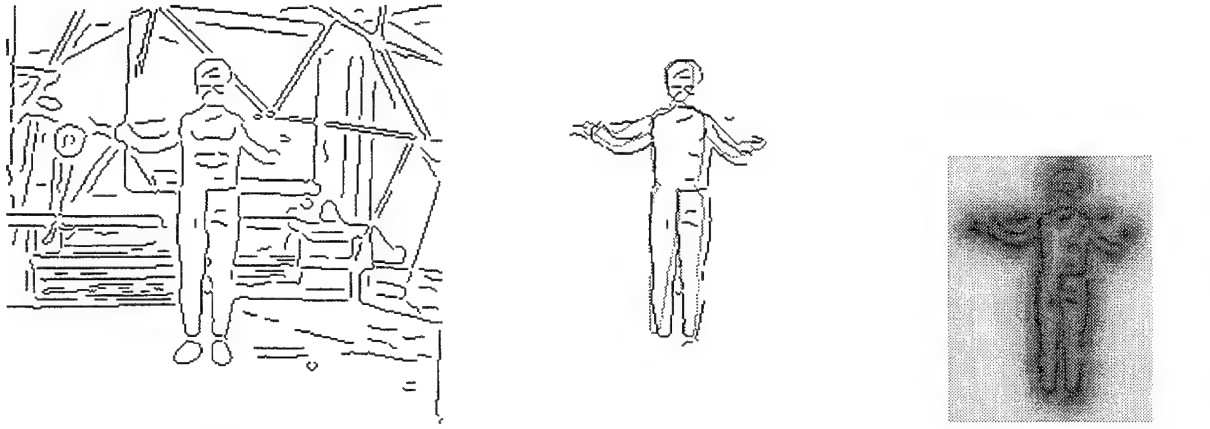
Its normalized version is

$$\overline{DD}(T, R) = DD(T, R)/|T| \tag{4}$$

$DD(T, R)$ can be efficiently obtained in a two-pass process by pre-computing the chamfer distance on a grid to the reference set. The resulting distance map is the so-called "chamfer image" (see Figures 4b and 4c). It would be efficient if we could use only $DD(M, S)$ during pose search (as done in [3]), where M and S are the projected model edges and scene edges, respectively. In that case, the scene chamfer image would have to be computed only once, followed by fast access for different model projections. However, using this measure alone has the disadvantage (which becomes apparent in experiments) that it does not contain information about how close the reference set is to the test set. For example, a single point can be really close to a large straight line, but we may not want to consider the two entities

very similar. We therefore use the *undirected* normalized chamfer distance

$$\overline{D}(T,R) = (\overline{DD}(T,R) + \overline{DD}(R,T))/2 \tag{5}$$



(a) (b) (c)

Figure 4: (a) Scene edge image (after preprocessing); (b) filtered edge image (model prediction in grey, accepted edges in black); (c) chamfer image.

A further modification is to perform outlier rejection on the distribution $dd(t,R)$. Points t for which $dd(t,R) > \theta$ are rejected outright; the mean μ and standard deviation σ of the resulting distribution is used to reject points t for which $dd(t,R) > \mu + 2\sigma$.

Other measures which work directly on the scene image could (and have) been used to evaluate a hypothesized model pose: correlation (see [24] and [9]) and average contrast value along the model edges (a measure commonly used in the snake literature). The reason we opted for preprocessing the scene image (i.e. applying an edge detector) and chamfer matching is that it provides a gradual measure of similarity between two contours while having a long-range effect in image space. It is gradual since it is based on distance contributions of many points along both model and scene contours; as two identically contours are moved apart in image space the average closest distance between points increases gradually. This effect is noticeable over a range up to a threshold θ, in the absence of noise. The two factors, graduality and long-range effect, make (chamfer) distance mapping a suitable evaluation measure to guide a search process. Correlation and average contrast along a contour, on the other hand, typically provide strong peak responses but rapidly declining off-peak responses.

Multi-view approach

By using a multi-view approach we achieve tighter 3D pose recovery and tracking of the human body than by using one view only; body poses and movements that are ambiguous from one view can be disambiguated from another view. We synthesize appearances of the human model for all the available views, and evaluate the appropriateness of a 3D pose based on the similarity measures for the individual views (see Figure 3b). Currently, the contributions from the different views are weighed inversely proportionally to the distance between the human torso center and the camera plane (this uses some simplifying assumptions, among them orthogonal projection). We plan to include a weighting scheme which reasons locally (per body unit) about the reliability of the observations.

Search

Search techniques are used to prune the high dimensional pose parameter space (see also [20]). We currently use *best-first* search; we do this because a reasonable initial state can be provided by a prediction component during tracking or by a bootstrapping method at start-up. The use of a well-behaved similarity measure derived from multiple views, as discussed before, is likely to lead to a search landscape with fairly wide and pronounced maxima around the correct parameter values; this can be well detected by a local search technique such as best-first. Nevertheless, the fact remains that the search space is very large and high-dimensional (22 dimensions per human, in our case); this makes "straight-on" search daunting. The proposed solution to this is *search space decomposition.* Define the original N-dimensional search space Σ at time t as

$$\Sigma = \{\{p_1\} \times \cdots \times \{p_N\}\}, \qquad p_i = \hat{p}_i - \Delta_{1i}, \ldots, \hat{p}_i + \Delta_{2i}, \text{ step } \Delta_{3i} \tag{6}$$

where $\hat{\mathbf{P}} = (\hat{p}_1, \ldots, \hat{p}_N)$ is the state prediction for time t. We define the decomposed search space Σ^* as

$$\Sigma^* = (\Sigma_1, \Sigma_2) \tag{7}$$

$$\Sigma_1 = \{\{p_{i_1}\} \times \cdots \times \{p_{i_M}\} \times \{\hat{p}_{i_{M+1}}\} \times \cdots \times \{\hat{p}_{i_N}\}\} \tag{8}$$

$$\Sigma_2 = \{\{\tilde{p}_{i_1}\} \times \cdots \times \{\tilde{p}_{i_M}\} \times \{p_{i_{M+1}}\} \times \cdots \times \{p_N\}\} \tag{9}$$

where $(\tilde{p}_{i_1}, \ldots, \tilde{p}_{i_M})$ is derived from the best solution to searching for Σ_1. The above search space decomposition can be applied recursively and can be represented by a tree in which non-leaf nodes represent search spaces to be further decomposed and leaf nodes are search spaces to be actually processed. The recursive scheme we propose for the pose recovery of K humans is illustrated in Figure 5. In order to search for the pose of the i-th human in the scene we synthesize humans $1, \ldots, i-1$ with the best pose parameters found so far, and synthesize humans $i+1, \ldots, K$ with their predicted pose parameters. Next we search for the best torso/head configuration of the i-th human while keeping the limbs at their predicted values, etc.

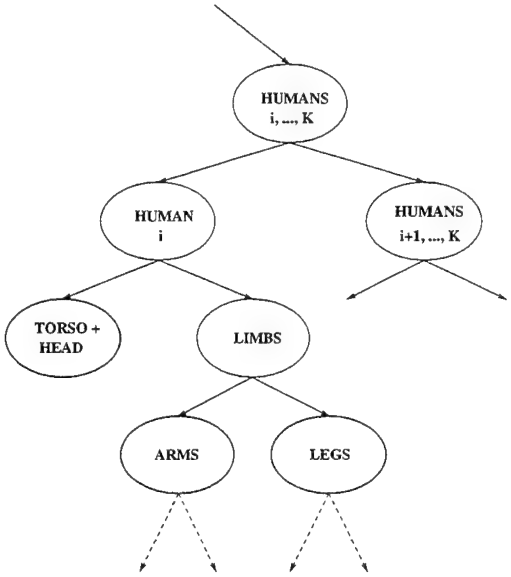


Figure 5: A decomposition of the pose-search space.

We have found in practice that it is more stable to include the torso-twist parameter in the arm (or leg) search space, instead of in the torso/head search space. This is because the observed contours of the torso alone are not very sensitive to twist. Given that we keep the root of the articulated figure fixed at the torso center, the dimensionalities of the search spaces we actually search are 5, 9, and 8, respectively.

Initialization

Our bootstrapping procedure for starting the tracking currently handles the case where the moving objects (i.e. humans) do not overlap and are positioned against a stationary

background. The procedure starts with background subtraction, followed by a thresholding operation to determine the region of interest; see Figure 6. This operation can be quite noisy, as shown in the figure. The aim is to determine from this binary image the major axis of the region of interest; in practice this is the axis of the prevalent torso-head configuration. Together with the major axis of another view, this allows the determination of the major 3D axis of the torso. Additional constraints regarding the position of the head along the axis (currently, implemented as a simple histogram technique) allow a fairly precise estimation of all torso parameters, with the exception of the torso twist which is searched for, together with the arm/leg parameters, in a coarse to fine fashion.

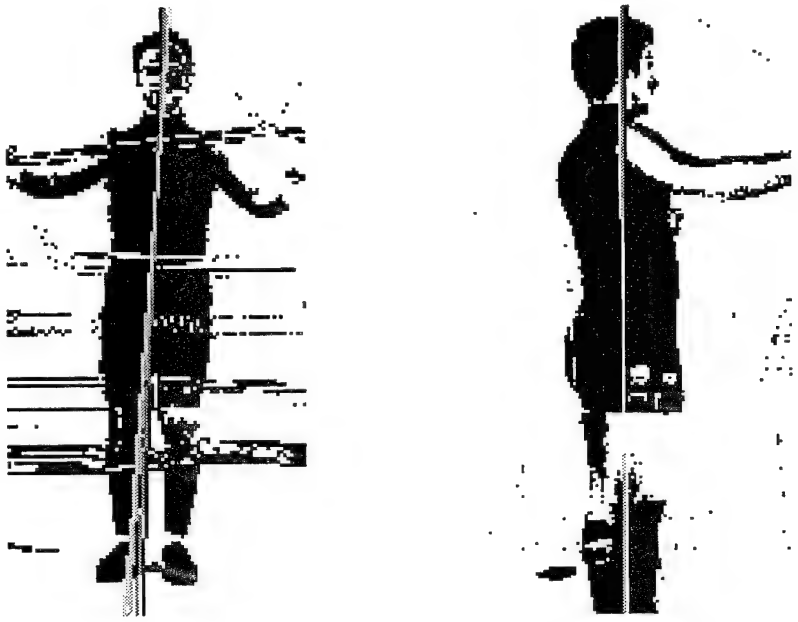


Figure 6: Robust major axis estimation using iterative PCA (cameras FRONT and RIGHT). Successive approximations to the major axis are shown in lighter colors.

The determination of the major axis can be achieved robustly by iteratively applying a principal component analysis (PCA) [14] on data points sampled from the region of interest. At each iteration the "best" major axis is computed using PCA and the distribution of the distances from the data points to this axis is computed. Data points whose distances to the current major axis are more than the mean plus twice the standard deviation are considered outliers and removed from the data set. This process results in the removal of the data points corresponding to the hands if they are located lateral to the torso, and also of other types of noise. The iterations are halted if the parameters of the major axis vary

by less than a user-defined fraction from one iteration to another. In Figure 6 the successive approximations to the major axis are shown by straight lines in increasingly light colors.

4.2 The other components

Our prediction component works in batch mode and uses a constant acceleration model for the pose parameters. In other words, a second-degree polynomial is fitted at times $t, \ldots, t-T+1$, and its extrapolated value at times $t+1$ is used for prediction. The synthesis component uses a standard graphics renderer to give the model projections for the various camera views. Finally, the image analysis component applies an edge detector to the real images, performs linking, and groups the edges into constant-curvature segments. These segments are each considered as a unit and either accepted into or rejected from the filtered scene edge map, a decision which is based on their directed chamfer distances to the projected model edges; see Figure 4. This process facilitates the removal of unwanted contours which could disturb the scene chamfer image (in Figure 4, for example, background edges around the head area in the original edge image are absent in the filtered edge image).

5 Experiments

We compiled a large data base containing multi-view images of human subjects involved in a variety of activities. These activities are of various degrees of complexity, ranging from single-person hand waving to the challenging two-person close interaction of the Argentine tango. The data was taken from four (near-) orthogonal views (FRONT, RIGHT, BACK and LEFT) with the cameras placed wide apart in the corners of a room for maximum coverage; see Figure 7. The background is fairly complex; many regions contain bar-like structures, and some regions are highly textured (observe the two VCR racks in the lower-right image of Figure 7). The subjects wore tight-fitting clothes. Their sleeves were of contrasting colors, simplifying the edge detection somewhat in cases where one body part occludes another.

Because of disk space and speed limitations, the more than one hour's worth of image data was first stored on (SVHS) video tape. A subset of this data was digitized (properly aligned by its time code (TC)), and makes up the HIA database, which currently contains

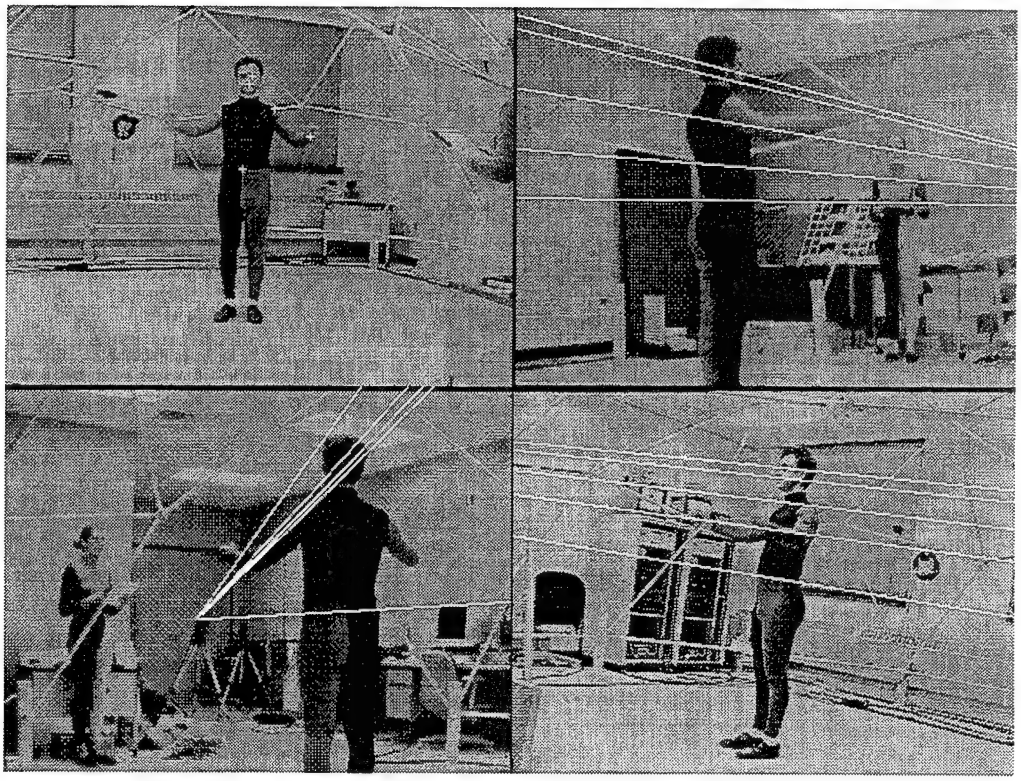


Figure 7: Epipolar geometry of cameras FRONT (upper-left), RIGHT (upper-right), BACK (lower-left) and LEFT (lower-right): epipolar lines are shown corresponding to the selected points from the view of camera FRONT.

more than 2500 frames in each of the four views.

The cameras were calibrated in a two-step process, first for the intrinsic parameters (individually) and then for the extrinsic parameters (in pairs). We used an iterative non-linear least square method to do this; it was developed by Szeliski and Kang [27] who kindly made it available to us. Figure 7 illustrates the outcome; the epipolar lines shown in the RIGHT, BACK and LEFT views correspond to the selected points in the FRONT view. One can see that corresponding points lie very close to or on top of the epipolar lines. Observe how all the epipolar lines emanate from one single point in the BACK view: the FRONT camera center lies within its view.

Our system is implemented under A.V.S. (Advanced Visualization System). Following its data flow network model, it consists of independently running modules, receiving and passing data through their interconnections. The implemented A.V.S. network bears a close resemblance to Figure 3. The parameter space was bounded in each angular dimension by

± 15 degrees, and in each xyz-dimension by ± 10 cm around the predicted parameter values. The discretization was 5 degrees and 5 cm, respectively. We kept these values constant during tracking.

Figures 8–13 illustrate tracking for persons DARIU and ELLEN. The movement performed can be described as raising the arms sideways to a 90 degree extension, followed by rotating both elbows forward. Moderate opposite torso movement takes place for balancing as the arms are moved forward and backwards. The current recovered 3D pose is illustrated by the projection of the model in the four views, shown in white. (The displayed model projections include for visual purposes the edges at the intersections of body parts; these were not included in the chamfer matching process.) It can be seen that tracking is quite successful, with a good fit for the recovered 3D pose of the model for the four views. Figure 14 shows some of the recovered pose parameters for the DARIU sequence. Figure 15 shows the result of movement recognition using a variant of Dynamic Time Warping (DTW), described in [6]; for the time-interval in which the elbows rotate forward, we use the left hand pose parameters derived from the ELLEN sequence as a template (see Figure 15a) and match them with the corresponding parameters of the DARIU sequence. Matching with DTW allows (limited) time-scale variations between patterns. The result is given in Figure 15b, where the DTW dissimilarity measure drops to a minimum when the corresponding pose pattern is detected in the DARIU sequence.

6 Discussion

As we process more sequences of our HIA database our aim is to be able to process the more complex sequences, involving fast-varying poses, multiple bodies and close interactions. One such example is the "Basico" sequence, in which two persons dance the basic steps of the Argentine tango at normal speed; see Figure 16. We show a manual positioning of the 3D models of the dancers.

We consider several improvements to our system. On the image processing level, we are interested in a tighter coupling between prediction and segmentation. Currently, the image processing component applies a general-purpose edge detector and uses prediction only

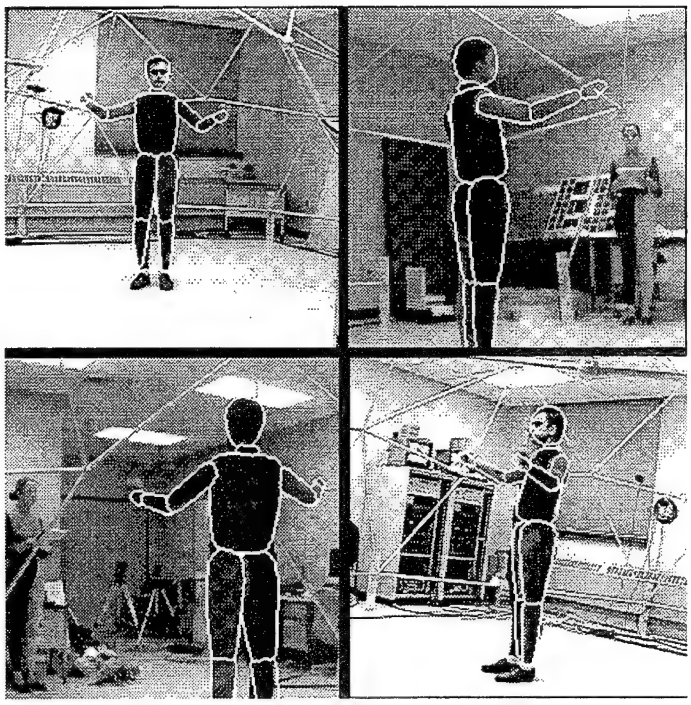


Figure 8: Tracking sequence D-TwoElbowRot, $t = 0$, cameras FRONT, RIGHT, BACK and LEFT.

for filtering purposes. We are interested in more actively using the prediction information through the use of deformable templates. On the algorithmic level, we are interested in methods of further constraining the search space, based on either image flow or stereo correspondence. Finally, for performance, we plan a parallel and distributed implementation of our system, an extension which is well supported by our approach and A.V.S.

7 Conclusions

We have presented a new vision system for the 3D model-based tracking of unconstrained human movement from multiple views. A large Humans-In-Action database has been compiled for which initial tracking results were shown. We can draw two conclusions from these initial experimental results. First, our calibration and human modeling procedures support a (perhaps surprisingly) good 3D localization of the model such that its projections match the all-around camera views. This is good news for the feasibility of *any* multi-view 3D model-based tracking method, not just ours. Second, the proposed pose recovery and tracking method based on, among others, the chamfer distance similarity measure, is indeed

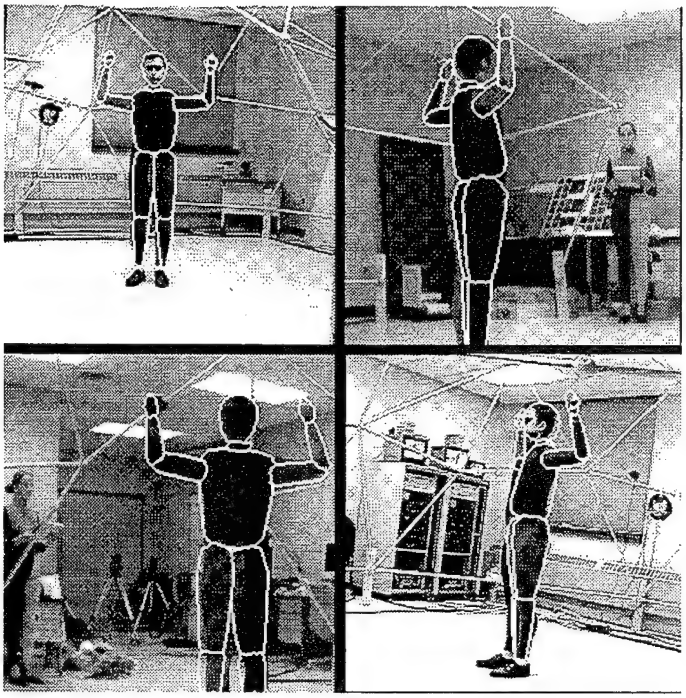


Figure 9: Tracking sequence D-TwoElbowRot: $t = 10$ (cameras FRONT, RIGHT, BACK and LEFT).

able to maintain a good fit over time. This is encouraging as we turn to the more complex sequences.

8 Acknowledgements

We would like to thank Ellen Koopmans, P.J. Narayanan and Pete Rander for their support in acquiring the Humans-In-Action database at CMU's 3D Studio.

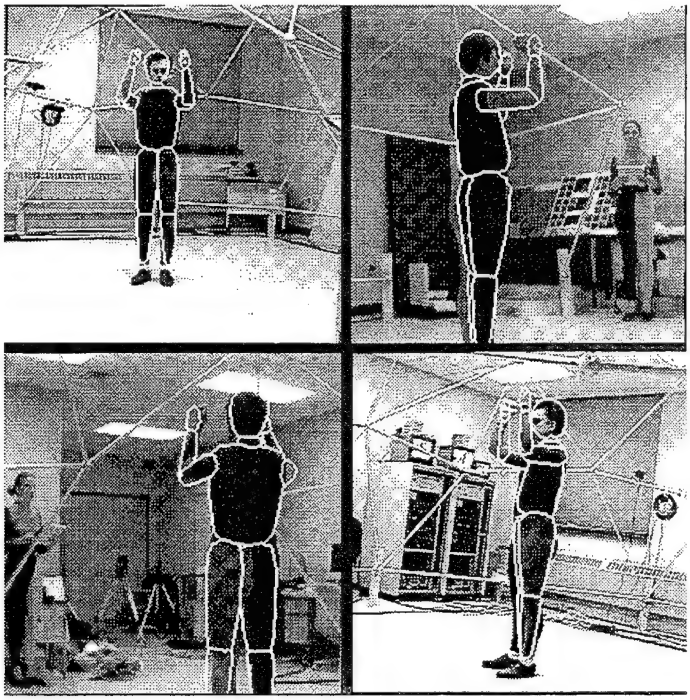


Figure 10: Tracking sequence D-TwoElbowRot: $t = 25$ (cameras FRONT, RIGHT, BACK and LEFT).

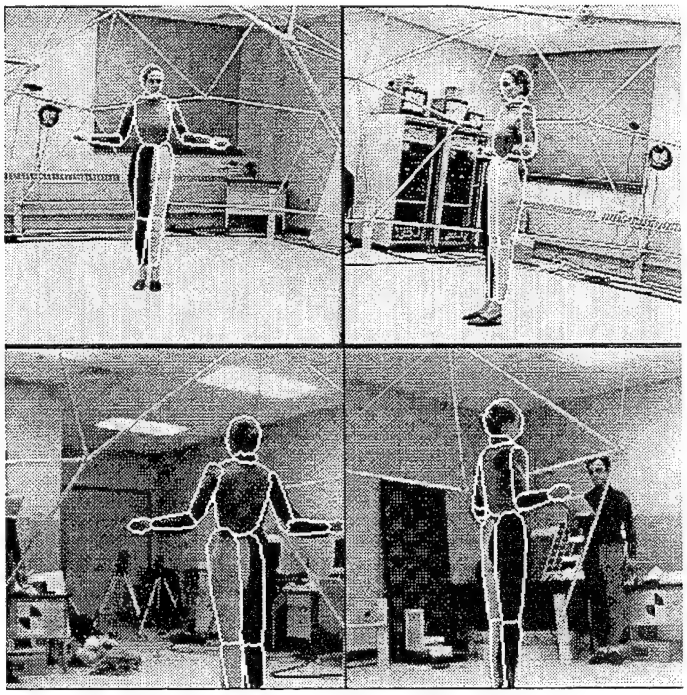


Figure 11: Tracking sequence E-TwoElbowRot: $t = 0$ (cameras FRONT, LEFT, BACK and RIGHT).

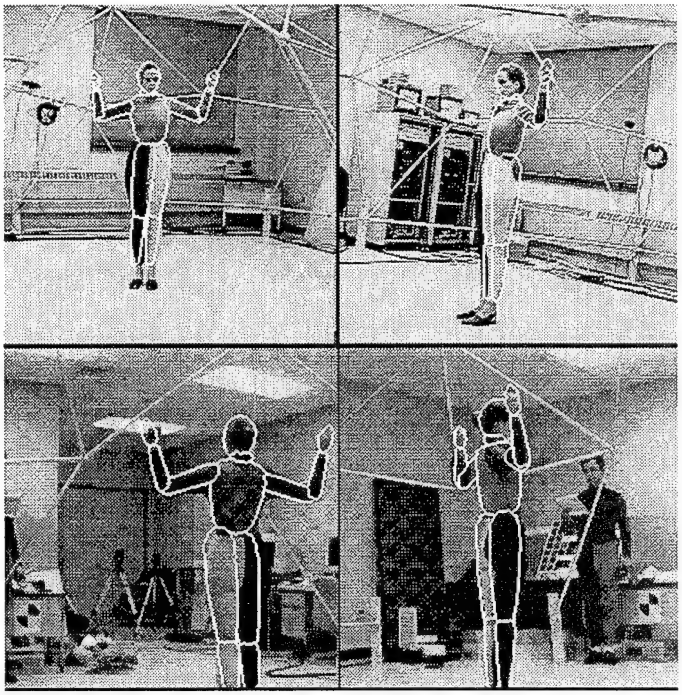


Figure 12: Tracking sequence E-TwoElbowRot: $t = 10$ (cameras FRONT, LEFT, BACK and RIGHT).

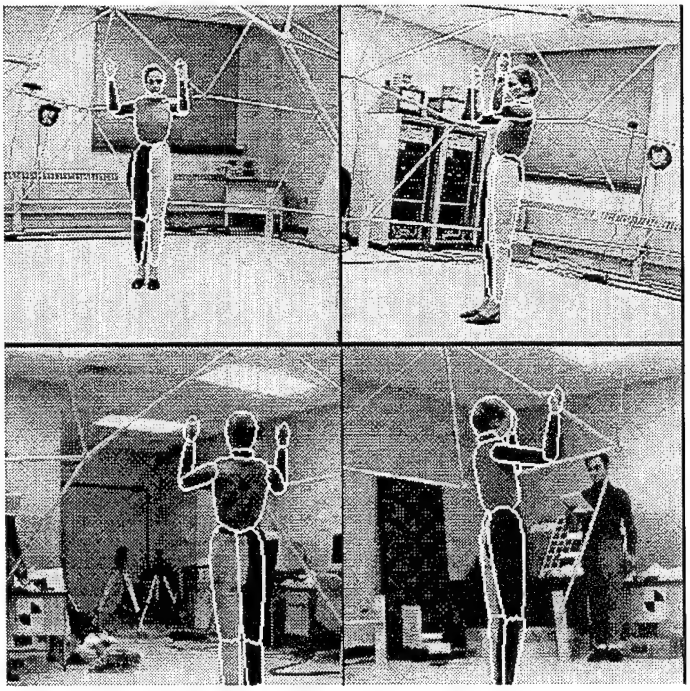


Figure 13: Tracking sequence E-TwoElbowRot: $t = 25$ (cameras FRONT, LEFT, BACK and RIGHT).

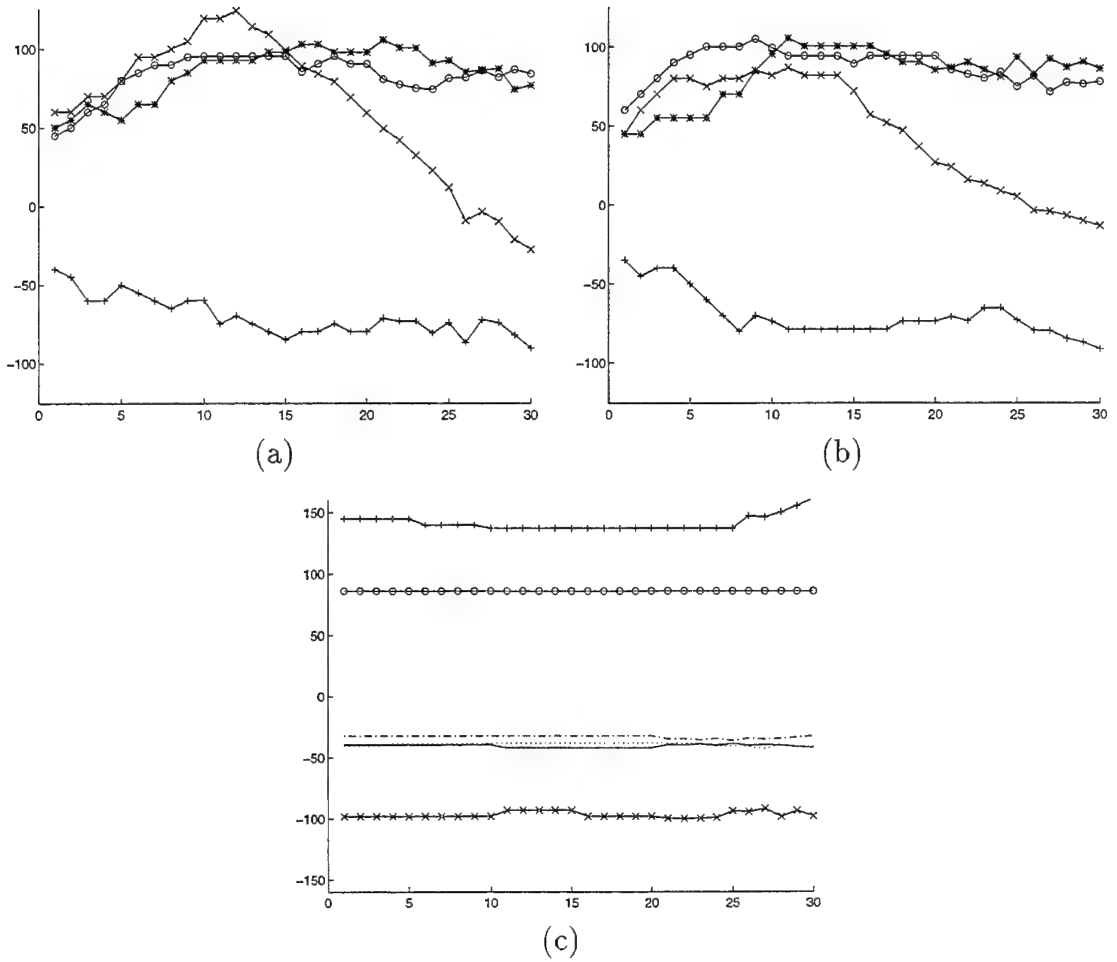


Figure 14: Recovered 3D pose parameters vs. frame number, D-TwoElbowRot; (a) and (b): LEFT and RIGHT ARM, abduction (x), elevation (o), twist (+) and extension angle (*) (c): TORSO, abduction (x), elevation (o), twist angle (+) and x- (dot), y- (dashdot), and z-coordinates (solid).

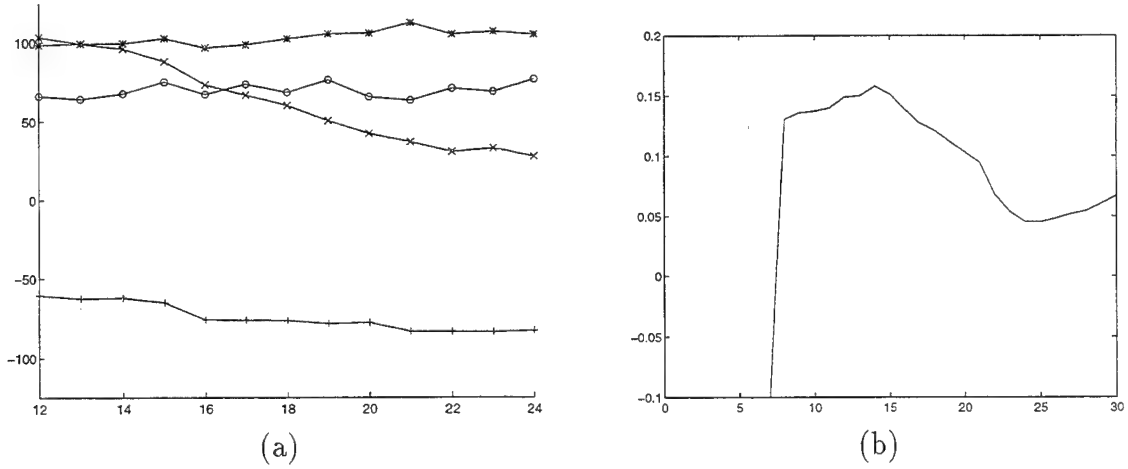


Figure 15: (a) A template T for the left arm movement, extracted from E-TwoElbowRot; (b) DTW dissimilarity measure of matching template T with the LEFT ARM pose parameters of D-TwoElbowRot.

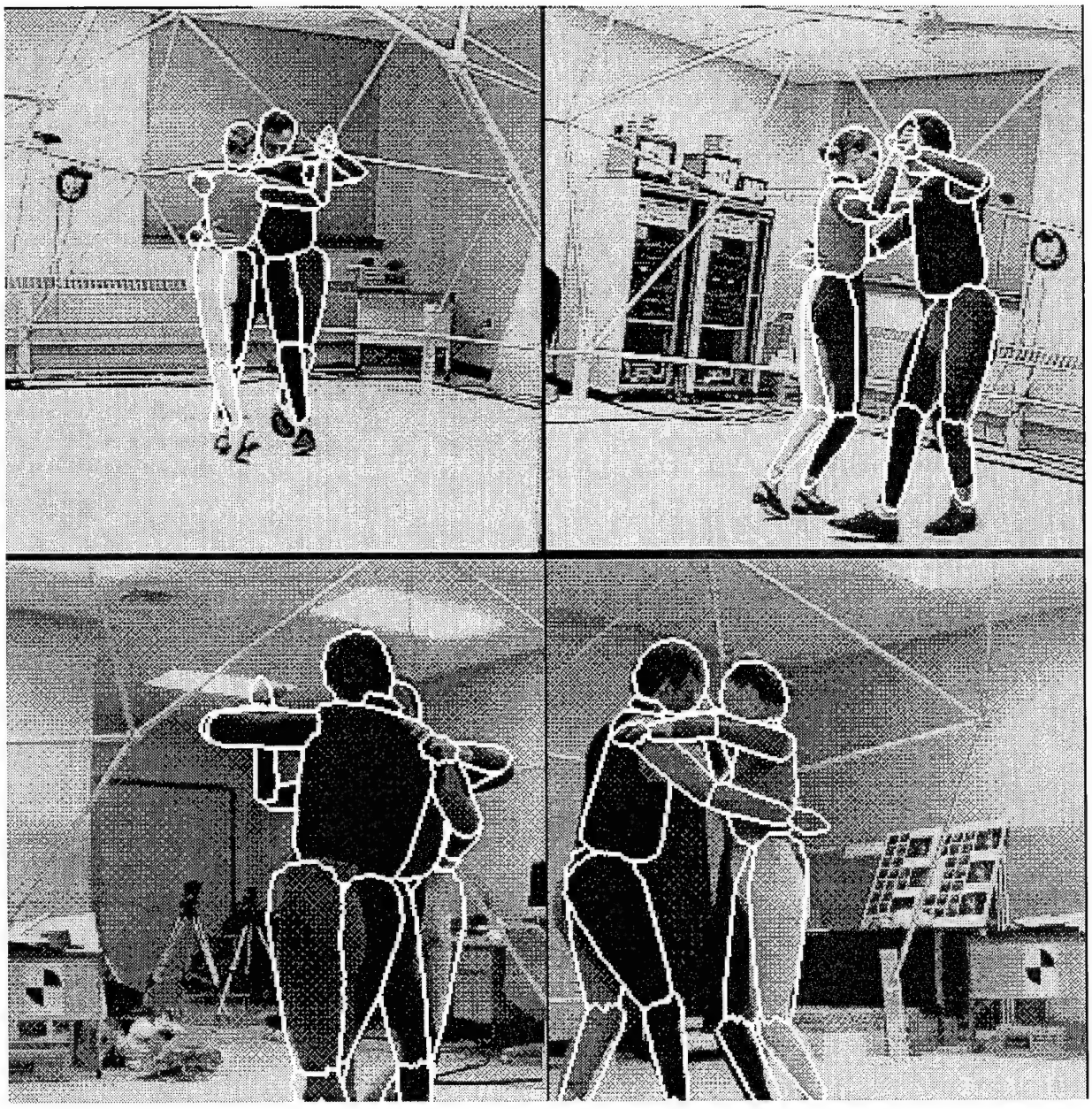


Figure 16: "Manual" 3D pose recovery for a couple dancing the Argentine tango (cameras FRONT, RIGHT, BACK and LEFT).

References

[1] K. Akita, "Image Sequence Analysis of Real World Human Motion," *Pattern Recognition*, vol. 17, pp. 73–83, 1984.

[2] N.I. Badler, C.B. Phillips, and B.L. Webber, *Simulating Humans*, Oxford University Press, Oxford, UK, 1993.

[3] H.G. Barrow et al., "Parametric Correspondence and Chamfer Matching: Two New Techniques For Image Matching," *Proc. IJCAI*, pp. 659–663, 1977.

[4] T. Darrell and A. Pentland, "Space-Time Gestures," *Looking at People, Proc. IJCAI*, 1993.

[5] A.C. Downton and H. Drouet, "Model-Based Image Analysis for Unconstrained Upper-Body Motion," *Proc. Intl. IEE Conf. on Image Processing and its Applications*, pp. 274–277, 1992.

[6] D.M. Gavrila and L.S. Davis, "Towards 3-D Model-based Tracking and Recognition of Human Movement," *Proc. Intl. Workshop on Face and Gesture Recognition*, Zurich, Switzerland, 1995.

[7] D.M. Gavrila and L.S. Davis, "3-D Model-based Tracking of Human Upper-body Movement: A Multi-view Approach," *Proc. IEEE Intl. Symp. on Computer Vision*, Coral Gables, FL, pp. 253–258, 1995.

[8] N. Goddard, "Incremental Model-Based Discrimination of Articulated Movement Direct from Motion Features," *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, 1994.

[9] L. Goncalves et al., "Monocular Tracking of the Human Arm in 3D," *Proc. ICCV*, pp. 764–770, 1995.

[10] Y. Guo, G. Xu and S. Tsuji, "Understanding Human Motion Patterns," *Proc. ICPR*, pp. 325–329, 1994.

[11] D. Hogg, "Model Based Vision: A Program to See a Walking Person," *Image and Vision Computing*, vol. 1, pp. 5–20, 1983.

[12] R. Holt, T.S. Huang, A. Netravali and R. Qian, "Determining Articulated Motion from Perspective Views: A Decomposition Approach," *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, 1994.

[13] S.S. Intille and A.F. Bobick, "Closed-World Tracking," *Proc. ICCV*, pp. 672–678, 1995.

[14] I.T. Jolliffe, *Principal Component Analysis*, Springer Verlag, New York, 1986.

[15] I. Kakadiaris and D. Metaxas, "3D Human Body Model Acquisition from Multiple Views," *Proc. ICCV*, pp. 618–623, 1995.

[16] M.K. Leung and Y.H. Yang, "First Sight: A Human Body Outline Labeling System," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 359–377, 1995.

[17] D. Marr and H.K. Nishihara, "Representation and Recognition of the Spatial Organization of Three Dimensional Shapes," *Proc. Royal Soc. London B*, vol. 200, pp. 269–294, 1978.

[18] D. Metaxas and D. Terzopoulos, "Shape and Nonrigid Motion Estimation through Physics-Based Synthesis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 580–591, 1993.

[19] S.A. Niyogi and E.H. Adelson, "Analyzing and Recognizing Walking figures in XYT," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 469–474, 1994.

[20] J. Ohya and F. Kishino, "Human Posture Estimation from Multiple Images Using Genetic Algorithm," *Proc. ICPR*, pp. 750–753, 1994.

[21] A. Pentland, "Automatic Extraction of Deformable Models," *Intl. J. Computer Vision*, vol. 4, pp. 107–126, 1990.

[22] F.J. Perales and J. Torres, "A System for Human Motion Matching between Synthetic and Real Images Based on a Biomechanic Graphical Model," *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, 1994.

[23] R. Polana and R. Nelson, "Low Level Recognition of Human Motion," *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, 1994.

[24] J. Rehg and T. Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects," *Proc. ICCV*, pp. 612–617, 1995.

[25] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," *CVGIP: Image Understanding*, Vol. 59, pp. 94–115, 1994.

[26] J. O'Rourke and N.I. Badler, "Model-based Image Analysis of Human Motion using Constraint Propagation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 522–536, 1980.

[27] R. Szeliski and S.B. Kang, "Recovering 3D Shape and Motion from Image Streams Using Nonlinear Least Squares," *J. Visual Communication and Image Representation*, vol. 5, pp. 10–28, 1994.

[28] J. Yamato, J. Ohya and K. Ishii, "Recognizing Human Action in Time-Sequential Images using Hidden Markov Model," *Proc. CVPR*, pp. 379–385, 1992.

[29] J. Zhao, *Moving Posture Reconstruction from Perspective Projections of Jointed Figure Motion*, Ph.D. Thesis, University of Pennsylvania, 1993.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1995	Technical Report

4. TITLE AND SUBTITLE
3D Model-Based Tracking of Humans in Action: A Multi-View Approach

5. FUNDING NUMBERS
N00014-95-1-0521

6. AUTHOR(S)
D.M. Gavrila and L.S. Davis

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer Vision Laboratory
Center for Automation Research
University of Maryland
College Park, MD 20742-3275

8. PERFORMING ORGANIZATION REPORT NUMBER
CAR-TR-799
CS-TR-3555

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Advanced Research Projects Agency
3701 N. Fairfax Drive, Arlington, VA 22203-1714
Office of Naval Research
800 North Quincy Street, Arlington, VA 22217-5660

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release.
Distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We present a vision system for the 3D model-based tracking of unconstrained human movement. Using image sequences acquired simultaneously from multiple views, we recover the 3D body pose at each time instant without the use of markers. The pose-recovery problem is formulated as a search problem and entails finding the pose parameters of a graphical human model whose synthesized appearance is most similar to the actual appearance of the real human in the multi-view images. The models used for this purpose are acquired from the images. We use a decomposition approach and a best-first technique to search through the high dimensional pose parameter space. A robust variant of chamfer matching is used as a fast similarity measure between synthesized and real edge images. We present initial tracking results from a large new Humans-In-Action (HIA) database containing more than 2500 frames in each of four orthogonal views. The four image streams are synchronized. They contain subjects involved in a variety of activities, of various degrees of complexity, ranging from simple one-person hand waving to two-person close interaction in the Argentine tango.

14. SUBJECT TERMS
Human body modeling, human body tracking, pose recovery

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- **DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank.
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.